

Annotated Image to Text and Speech Synthesis using Cloud Application Programming Interfaces

A.NeelaMadheswari*¹, R.S Prathiksaa, A.Reeja Mary, M.V.SnehaPriya

Computer Science Engineering, Mahendra Engineering College, Mahendhirapuri,
Mallasamudram - 637503, Tamil Nadu, India.

ABSTRACT: A number of applications emerging related to image to text conversion or image to speech conversion which helps us in various ways. This paper focussed on the conversion of Image-to-Speech in two different phases. The supporting tools used to implement this work are tesseract-OCR, machine learning based libraries, and speech recognizer using Python. For a given image the texted regions are partitioned using morphological operations. An annotated image to text and speech synthesis using cloud API is done. This work focussed on the conversion of input image into text, and further the text is converted into English speech and then translated into speech of various Indian languages. The proposed system is useful for visually impaired persons who are not able to read or recognize the text given in many sign boards or images, and also in many more applications where we are in need of Image-to-Speech synthesis. It also focussed on translating the English speech into a number of Indian languages such as Nepali, Urdu, Tamil, Malayalam, Bengali, Gujarati, Hindi, Kannada, and Marathi.

KEYWORDS: OCR, text-to-speech, image-to-text, execution time, image file format

<https://doi.org/10.29294/IJASE.9.1.2022.2511-2516> ©2022 Mahendrapublications.com, All rights reserved

INTRODUCTION

An image contains perceptual contents and semantic contents. Perceptual content focus on colours, shapes, textures, intensities and the semantic content focus on objects and their relationships. Text is the major content that comes under semantic content. For content analysis, text extraction from images is very significant. A number of applications involved in text extraction are vehicle license plate recognition, automatic bank check processing, signboard detection and translation, assisting visually impaired persons etc [1].

Alakbar et al., [2] in their proposed work, developed and evaluated the speech synthesis system based on deep learning models for Azerbaijani language. Their system is used to analyse text-to-speech synthesis and using Azerbaijani language. Mark et al., [3], proposed Image2speech, a system for automatic generation of audio description of images. But their system cause many errors due to misrecognition of objects and actions in the image. As though having a few errors, their system generated the spoken description of image directly without first generating text.

A multimodal information bottleneck approach is proposed by Shuang et al., [4]. Their model aims to translate one modality to another

by skipping an intermediate modality shared by two different datasets. They focussed on two main perspectives: i) image-to-speech synthesis, and ii) effectiveness of multimodal modelling. Image-text samples and audio-text samples are considered and concluded that better results obtained for image-to-speech synthesis.

Anindya and Sriparna [5] proposed self-supervised deep learning based approach for image to text and text to image generations. They used StackGAN-based autoencoder model and also LSTM based text-autoencoder. Their work focussed on generation of text in English language. Wei et al, [6] proposed a model for direct synthesis of speech for images. They conducted experiments using MSCOCO dataset and used English language for speech synthesis using Learned Segmental units.

Ifeanyi et al., [7] proposed a novel adaptive binarization method based on wavelet filter is proposed. This approach was processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. They evaluated their adaptive method on complex scene images of ICDAR 2005 database. Arora & Shetty [8] reported some problems in text recognition and retrieval. Automated

*Corresponding Author: neelamadheswaria@mahendra.info

Received: 10.05.2022

Accepted: 27.06.2022

Published on: 01.08.2022

NeelaMadheswari et al.,

optical character recognition (OCR) tools do not supply a complete solution and in most cases human inspection is required. They suggest a novel text recognition algorithm based on usage of fuzzy logic rules relying on statistical data of the analysed font. The new approach combines letter statistics and correlation coefficients in a set of fuzzy based rules, enabling the recognition of distorted letters that may not be retrieved otherwise. They focussed on rashi fonts associated with commentaries of the bible that are actually handwritten calligraphy.

Shrivastava [9] proposed a recognition scheme for the Indian script of devanagari. They used approach does not require word to character segmentation, which is one of the most common reasons for high word error rate. They reported a reduction of more than 20% in word error rate and over 9% reduction in character error rate while comparing with the best available OCR system. Douglas et.al, [10] proposed an architectural paradigm for the text-to-image conversion method using neural architecture. They claimed that their proposed work is the direct conversion of text-to-image in a single stage.

Images play a vital role in order to reach the thought or idea to humans much better than the text. But it is not possible to conclude the same for various situations like the people who are visually impaired cannot able to identify these images to gather the given information since there is no standard way of conveying images to them. In this situation, if the given image is converted into speech, it is very fruitful for the visually challenged people. The proposed work in this paper is focusing on this objective to make the contents in the images to be reached to those people who are visually challenged by converting the text from the given images into speech. Also the given speech is translated into a number of Indian languages such as Nepali, Urdu, Tamil, Malayalam, Bengali, Gujarati, Hindi, Kannada, and Marathi.

2. SYSTEM METHODOLOGY

The proposed system comprise of two phases. The first phase utilizes an English annotated image as input and converted into a text in txt file format using English language. In the second phase, the output of the first phase is considered as input that is the converted English text is given as input and is converted into speech in the form of mp3 audio file using English language. Then the speech in English is translated to a number of Indian languages.

Every time the input image is translated according to our requirements and the translated speech is also saved automatically. Similarly audio file for English speech is translated to Indian language and saved in the form of mp3 sound file. The output audio file size is noted for every execution of the native languages translated. The system is implemented using Python. Pyttsx3 is imported and utilized. The proposed system model for image to speech synthesis is given in figure 1. Frequency is calculated using the formula 1 as below [11]:

$$\text{Frequency} = \text{bit rate} / (\text{bit depth} \times \text{channels}) \quad (1)$$

The proposed system is analysed using various parameters: i) various image file formats for input image, ii) execution time for the conversion of annotated image to text output in English language, iii) execution time for the conversion of converted English text into speech of various Indian languages, iv) memory size of the given input image file, v) memory size of the audio file generated in various Indian languages, vi) frequency of the audio file. The code for running the proposed system is given in Figure 3, and the auto saved mp3 file format while running the proposed system is shown in Figure 4.

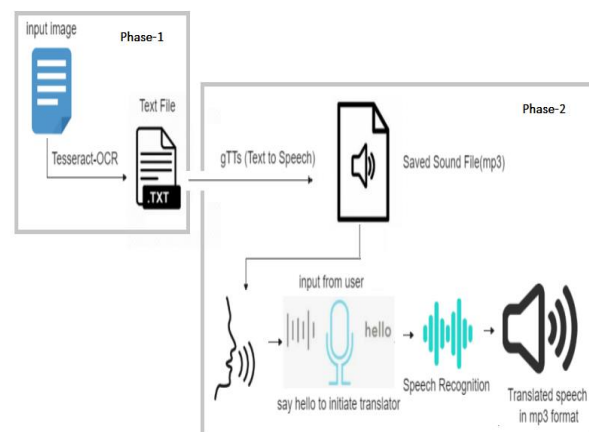


Figure 1: Image-to-Speech Synthesis - Proposed System model

3. IMAGE-TO-TEXT SYNTHESIS

The first phase of the proposed system is Image-to-Text synthesis using pytesseract, an OCR tool in Python in order to read and recognize text in Images. The input annotated image considered for the proposed system is given in Figure 2 [12]. This file is in jpg file format. The same image file is converted into

bmp file format and png file format for the same dimensions and the corresponding memory size of those image files are noted. The text file output is in the txt file format and the text in

that file is: "Once you replace negative thoughts with positive ones, you'll start having positive results." Willie Nelson.

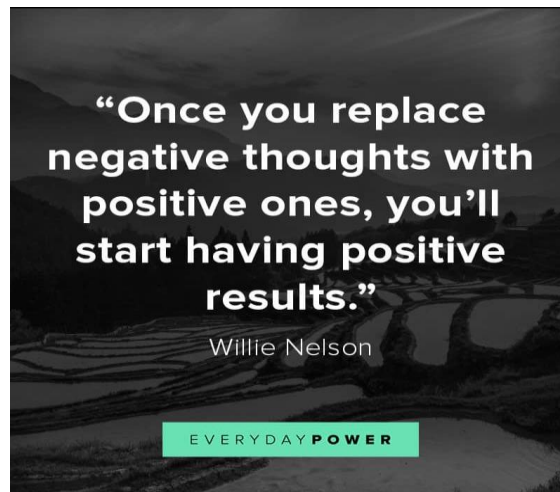


Figure 2: Annotated Input Image

Table 1: Image to Text Synthesis

| S.No | Image file type | Dimension | Image size (bit depth) | Image size (in KB) | Execution time (in ms) |
|------|-----------------|-----------|------------------------|--------------------|------------------------|
| 1 | jpg | 800 x 800 | 24 | 6 | 0.005 |
| 2 | bmp | 800 x 800 | 32 | 2.501 | 0.005 |
| 3 | png | 800 x 800 | 8 | 156 | 0.017 |

Table 1 specifies the image dimensions, image size in bit depth, actual image size in memory in terms of kilo bytes, execution time for the conversion of image to text. From Table 1, it is clearly observed that bmp file occupies less memory when compared to jpg and png file formats. While considering the execution time for the conversion of image into text, bmp and jpg files gives the conversion very faster when compare to png file format. It is also observed that while converting the jpg file to bmp file, the bit depth of the image also changes from 24 bit to 32 bit, and while converting the jpg file to png file, the bit depth of the image changes from 24 bit to 8 bit of depth.

1. TEXT-TO-SPEECH SYNTHESIS FOR INDIAN LANGUAGES

The output of the Image-to-text synthesis is considered as an input for this text-to-speech synthesis process. The input text file in English language is converted into speech in English language of mp3 file format using gTTS API support of Python. In order to perform the translation from English speech to various Indian languages, user has to initiate the translator by saying any word, here in the

system, hello is said. After this initiation, the given English speech is converted into our required Indian language speech and the output is saved in mp3 file format.

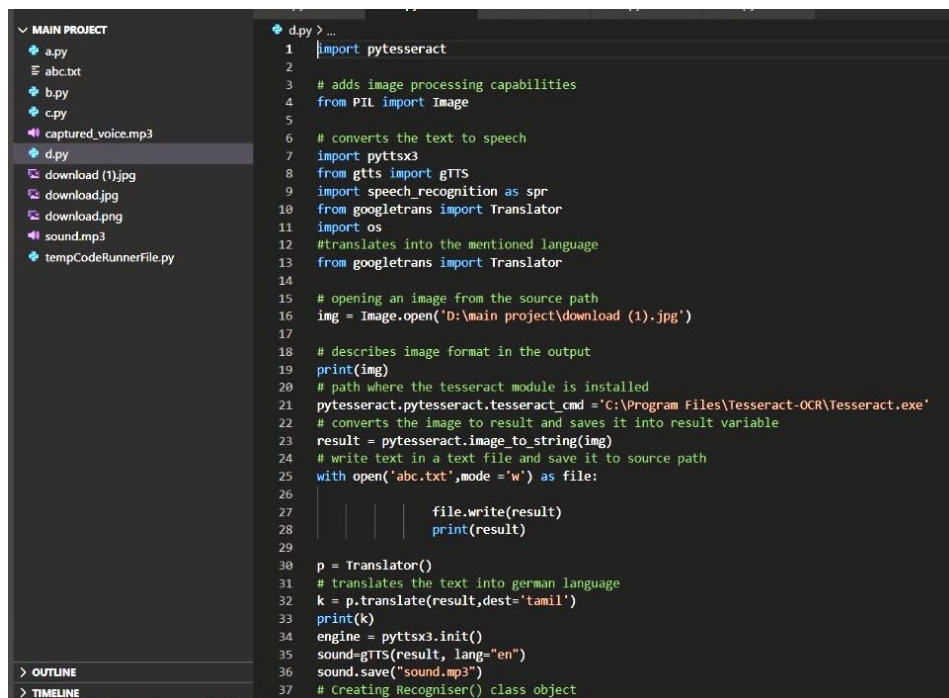
The proposed system supports up to nine Indian languages namely: i) Nepali, ii) Urdu, iii) Tamil, iv) Malayalam, v) Bengali, vi) Gujarati, vii) Kannada, viii) Hindi, and ix) Marathi. While converting the given text, various parameters are calculated such as i) execution time for the conversion of text to speech for various Indian languages, ii) Audio bit rate for output file, iii) Audio file size for output file, iv) Frequency of the output file. It is given in Table 2.

As per the results mentioned in Table 2, it is observed that Malayalam takes much amount of time for the conversion of text into speech in the case of jpg and bmp file formats whereas Marathi takes much amount of time for the conversion of text into speech in the case of png file format. In the case of audio bit rate, audio file size and frequency, it is more or less in the similar values for jpg files for all Indian languages, bmp files for all Indian languages and png files for all Indian languages.

NeelaMadheswari et al.,

Table 2: Text-to-Speech Synthesis for various Indian languages

| S.No | Image file format | Destination language | Execution time (in ms) | Audio bit rate (kbps) | Audio file size (Kb) | Frequency |
|------|-------------------|----------------------|------------------------|-----------------------|----------------------|-----------|
| 1 | jpg | Tamil | 0.0050 | 32 | 32 | 0.67 |
| 2 | jpg | Malayalam | 0.0080 | 32 | 32 | 0.67 |
| 3 | jpg | Bengali | 0.0010 | 32 | 32 | 0.67 |
| 4 | jpg | Gujarati | 0.0019 | 32 | 32 | 0.67 |
| 5 | jpg | Hindi | 0.0010 | 32 | 32 | 0.67 |
| 6 | jpg | Kannada | 0.0009 | 32 | 32 | 0.67 |
| 7 | jpg | Marathi | 0.0010 | 32 | 47 | 0.67 |
| 8 | jpg | Nepali | 0.0020 | 32 | 31 | 0.67 |
| 9 | jpg | Urdu | 0.0010 | 32 | 34 | 0.67 |
| 10 | bmp | Tamil | 0.0050 | 32 | 32 | 0.50 |
| 11 | bmp | Malayalam | 0.0067 | 32 | 26 | 0.50 |
| 12 | bmp | Bengali | 0.0010 | 32 | 32 | 0.50 |
| 13 | bmp | Gujarati | 0.0020 | 32 | 39 | 0.50 |
| 14 | bmp | Hindi | 0.0010 | 32 | 32 | 0.50 |
| 15 | bmp | Kannada | 0.0010 | 32 | 32 | 0.50 |
| 16 | bmp | Marathi | 0.0010 | 32 | 34 | 0.50 |
| 17 | bmp | Nepali | 0.0020 | 32 | 47 | 0.50 |
| 18 | bmp | Urdu | 0.0010 | 32 | 31 | 0.50 |
| 19 | png | Tamil | 0.0050 | 32 | 30 | 2.00 |
| 20 | png | Malayalam | 0.0029 | 32 | 26 | 2.00 |
| 21 | png | Bengali | 0.0055 | 32 | 32 | 2.00 |
| 22 | png | Gujarati | 0.0010 | 32 | 39 | 2.00 |
| 23 | png | Hindi | 0.0245 | 32 | 32 | 2.00 |
| 24 | png | Kannada | 0.0149 | 32 | 32 | 2.00 |
| 25 | png | Marathi | 0.0595 | 32 | 32 | 2.00 |
| 26 | png | Nepali | 0.0010 | 32 | 47 | 2.00 |
| 27 | png | Urdu | 0.0027 | 32 | 31 | 2.00 |



```

1 import pytesseract
2
3 # adds image processing capabilities
4 from PIL import Image
5
6 # converts the text to speech
7 import pyttsx3
8 from gtts import gTTS
9 import speech_recognition as sr
10 from googletrans import Translator
11 import os
12 # translates into the mentioned language
13 from googletrans import Translator
14
15 # opening an image from the source path
16 img = Image.open('D:\main project\download (1).jpg')
17
18 # describes image format in the output
19 print(img)
20 # path where the tesseract module is installed
21 pytesseract.pytesseract.tesseract_cmd = 'c:\Program Files\Tesseract-OCR\Tesseract.exe'
22 # converts the image to result and saves it into result variable
23 result = pytesseract.image_to_string(img)
24 # write text in a text file and save it to source path
25 with open('abc.txt', mode='w') as file:
26     file.write(result)
27     print(result)
28
29 p = Translator()
30 # translates the text into german language
31 k = p.translate(result, dest='tamil')
32 print(k)
33 engine = pyttsx3.init()
34 sound = gTTS(result, lang='en')
35 sound.save("sound.mp3")
36 # Creating Recogniser() class object

```

Figure 3: Python code in Visual Studio environment

NeelaMadheswari et al.,

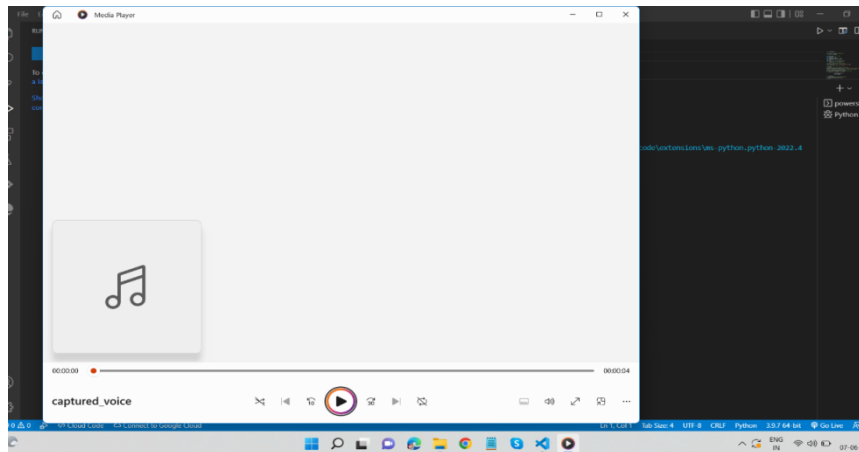


Figure 4: Auto saved mp3 file format

CONCLUSION

The proposed system gives an overview of the conversion of an annotated image file into text and further the text is converted into speech of various Indian languages using Python. The observed results showed that conversion of jpg file to png file takes much execution time for the conversion of image to text, while converting from text to speech, Malayalam language takes much execution time in the case of jpg and bmp file formats and Marathi takes much execution time for the conversion of text into speech while considering the input image as png file format. In the future, it can be extended to provide various languages images, text-to-speech synthesis from the language of any country to the language of any other countries. Further work is in progress, how to include emotion other than one language while synthesizing speech and also find the methods of improving quality of speech.

REFERENCES

- [1] Divyagera, Neelu Jain.2015. Comparison of Text Extraction Techniques – A Review, International Journal of Innovative Research in Computer and Communication Engineering, 3(2) Feb 621-626.
- [2] Alakbar Valizada., SevilJafarova., Emin Sultanov., Samir Rustamov. 2021. Development and Evaluation of Speech Synthesis System based on Deep Learning Models, Symmetry. 13(5), 819.
- [3] Mark-Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg and Francesco Ciannella. 2017. Image2speech: Automatically generating audio descriptions of images, Jelinek Speech and Language Technology Workshop.
- [4] Shuang Ma, Daniel Mc Duff and Yale Song. 2019. Unpaired Image-to-Speech Synthesis with Multimodal Information Bottleneck, Computer Vision and Pattern recognition, ICCV 2019, arXiv:1908.07094.
- [5] Anindya Sundar Das, Sriparna Saha. 2021. Self-Supervised Image-to-Text and Text-to-Image Synthesis, Computer Vision and Pattern Recognition, ICONIP 2021, arXiv:2112.04928v1.
- [6] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song and James Glass. 2020. Text-Free Image-to-Speech Synthesis using Learned Segmental Units, arXiv: 2012.15454.
- [7] Nwakanmalfeanyi, OhuigboIkenna and OkpalaIzunna. 2014. Text-To-Speech Synthesis, International Journal of Research in Information Technology, 2(5) May 154-163.
- [8] AnuArora, Anjali Shetty. 2014. Common Problems faced by Visually Impaired people, International Journal of Science and Research, 3(10) Oct 2002-2005.
- [9] Prashant Srivastava, Pradeep Kumar. 2015. Disability, Its Issues and Challenges: Psychosocial and Legal Aspects in Indian Scenario, Delhi Psychiatry Journal, 18(1), Apr 195-205.
- [10] Douglas M. Souza, Jonas Wehrmann, Duncan D.Ruiz. 2020. Efficient Neural Architecture for Text-to-Image Synthesis, arXiv.2004.11437v1.
- [11] Frequency formula, <https://www.bbc.co.uk/bitesize/guides/z>

- 7vc7ty/revision/4, accessed on 10.6.2022.
- [12] Annotated Input Image, <https://everydaypower.com/wp-content/uploads/2017/03/Inspirational-quote-13.jpg>, accessed on 1.4.2022.
- [13] Ranjit Ghosal., Ayan Banerjee. 2021. An Improved Scene Text and Document Image Binarization Scheme, 4th International Conference on Recent Advances in Information Technology (RAIT), 1-6.
- [14] Archana Balyan., Agrawal, S.S., Amita Dev. 2013. Speech Synthesis: A Review, International Journal of Engineering Research & Technology, 2(6).